EP3260: Machine Learning Over Networks
Computer Assignment 3
Due Date: March 7, 2023

Computer Assignment 3 - Training a neural network

Consider optimization problem

$$\underset{\mathbf{W}_1,\mathbf{W}_2,\mathbf{w}_3}{\text{minimize}} \frac{1}{N}\sum_{i\in[N]} \|\mathbf{w}_3\mathbf{s}(\mathbf{W}_2\mathbf{s}(\mathbf{W}_1\mathbf{x}_i) - \mathbf{y}_i\|_2^2,$$

where $\mathbf{s}(\mathbf{x}) = 1/(1+\exp(-\mathbf{x}))$. You may add your choice of regularizer. Using the "Individual household electric power consumption" and "Greenhouse Gas Observing Network" datasets, address the following questions:

(a) Try to solve this optimization task with proper choices of size of decision variables (matrix $\mathbf{W}_1$, matrix $\mathbf{W}_2$, and vector $\mathbf{w}_3$) using GD, perturbed GD, SGD, SVRG, and block coordinate descent. For the SGD method, you may use the mini-batch version.

(b) Compare these solvers in terms complexity of hyper-parameter tunning, convergence time, convergence rate (in terms of # outer-loop iterations), and memory requirement

- Adding regularizer :

minimize $\frac{1}{N}\sum_{i\in[N]} \|w_3 s(W_2 s(W_1 x_i)) - y_i\|_2^2 + \lambda\left(\|W_1\|_2^2 + \|W_2\|_2^2 + \|w_3\|_2^2\right)$
W_1, W_2, w_3

(b) According to the figures :

Hyper-parameter tunning :

GD, PGD, and BCD : only 1 hyper-parameter (learning rate) $\longrightarrow$ lowest complexity in terms of hyper-parameter tunning

SGD and SVRG : 2 hyper-parameters (learning rate, mini-batch size)

convergence time :

GD, PGD, and BCD : slower

SGD and SVRG : faster

convergence rate :

GD, PGD, and BCD : slower

SGD and SVRG : faster

SGD and SVRG update the weight matrices W_1 and W_2 and the vector w_3 more frequently

Memory requirement :

GD, PGD, and BCD : lowest memory requirement (only weight matrices and vector are stored)

SGD and SVRG : highest memory requirement

weight matrices & vector + mini-batch of training samples (in SGD) and control variates (in SVRG),

layer_0 $= x = f^0$

layer_1 $= s(w_i x_i) = s(w_1 f^{(0)}) = f^{(1)}$

layer_2 $= s(w_2 s(w_i x_i)) = s(w_2 f^{(1)}) = f^{(2)}$

layer_3 $= w_3 s(w_2 s(w_i x_i)) = w_3 f^{(2)} = f^{(3)}$

$f = \| \underbrace{f^{(3)} - y_i}_{\text{error}} \|^2 \longrightarrow \nabla f = 2 \cdot \text{error} \cdot \nabla f^3$

$\nabla f^{(3)}_{w_3} = f^{(2)}$

layer_3_delta $= 2 \times \text{error} \times f^2$ (∇ with respect to w_3)

layer_2_delta =

2 error $\times$ w_3 $\nabla f^{(2)}_{(w_2)}$ w.r.t. w_2

layer_1_delta =

2 error w_3 w_2 x_i $\nabla f^{(1)}_{w_1}$ w.r.t w_1